Econometric Theory, **19**, 2003, 923–943. Printed in the United States of America.
DOI: 10.1017/S0266466603196028

GENERALIZED EMPIRICAL LIKELIHOOD–BASED MODEL SELECTION CRITERIA FOR MOMENT CONDITION MODELS

HAN HONG AND BRUCE PRESTON
Princeton University

MATTHEW SHUM
Johns Hopkins University

This paper proposes model selection criteria (MSC) for unconditional moment models using generalized empirical likelihood (GEL) statistics. The use of GEL-statistics in lieu of *J*-statistics (in the spirit of Andrews, 1999, *Econometrica* 67, 543–564; and Andrews and Lu, 2001, *Journal of Econometrics* 101, 123–164) leads to an alternative interpretation of the MSCs that emphasizes the common information-theoretic rationale underlying model selection procedures for both parametric and semiparametric models. The result of this paper also provides a GEL-based model selection alternative to the information criteria–based nonnested tests for generalized method of moments models considered in Kitamura (2000, University of Wisconsin). The results of a Monte Carlo experiment are reported to illustrate the finite-sample performance of the selection criteria and their impact on parameter estimation.

1. INTRODUCTION

Exploiting insights from the recent literature on generalized empirical likelihood (GEL) estimation as an alternative to optimal generalized method of moments (GMM) estimation (cf. Qin and Lawless, 1994; Kitamura and Stutzer, 1997; Kitamura, 1997; Imbens, Spady, and Johnson, 1998; Ahn, Kitamura, and Tripathi, 2001; Newey and Smith, 2000; Smith, 1997), we propose model and moment selection criteria (MSC) for unconditional moment condition models based on the GEL statistic, in the spirit of Andrews (1999) and Andrews and Lu (2001). In these papers, Andrews and Lu investigate MSC for unconditional moment models using the GMM *J*-statistics (*J*-MSC). In this paper, we replace

The authors gratefully acknowledge support from the NSF (Hong: SES-0079495, Shum: SES-0003352) and the Fellowship of Woodrow Wilson Scholars (Preston). We thank the co-editor Don Andrews, Xiaohong Chen, John Geweke, Bo Honore, Yuichi Kitamura, Serena Ng, Harry Paarsch, Gautam Tripathi, and two anonymous referees for insightful suggestions and helpful comments. Address correspondence to: Han Hong, Department of Economics, Fisher Hall, Princeton University, Princeton, NJ 08544, USA; e-mail: doubleh@phoenix.Princeton.edu.

 0266-4666/03 $12.00

the *J*-statistics with the GEL-statistics in the construction of the MSCs. We also provide GEL analogs of the *J*-statistic-based "upward" and "downward" testing procedures considered in Andrews and Lu (2001).

As an example, let (b, c) denote a pair of model and moment selection vectors.[1] The GEL-MSC criterion selects the pair of vectors (b, c) that minimizes $GEL_n(b, c) - (|c| - |b|)\log n$, where GEL_n is the GEL function defined in the next section.

The use of GEL-statistics in lieu of *J*-statistics allows an alternative interpretation of the MSC and provides an information-theoretical analogy with MSCs in standard parametric likelihood models. Depending on the choice of the carrier function (defined subsequently), the GEL approach (see Newey and Smith, 2000; Smith, 1997) includes as special cases the empirical likelihood function (EL) of Qin and Lawless (1994), the exponential tilting function (ET) of Kitamura and Stutzer (1997), the Cressie–Read discrepancy statistics (CR) of Imbens, Spady, and Johnson (1998), and the continuous updating GMM function (CUE) of Hansen, Heaton, and Yaron (1996). For example, when we use EL-based MSC, our proposed selection criterion selects the model with the smallest Kullback-Leibler information criterion (KLIC) from the true underlying probability measure to the class of probability distributions implied by the moment conditions, and among those with the smallest KLIC it selects the one with the most parsimonious parameterization (or with the largest number of overidentification conditions).

The use of an information-theoretical approach for GMM model selection can be found in Kitamura (2000) and Ramalho and Smith (2002). Kitamura (2000) has developed information-theoretic nonparametric likelihood ratio tests to choose between nonnested moment condition models. Smith (1997) proposes nonnested Cox tests between GMM models using GEL functions. The results of this paper provide a GEL-based MSC alternative to the nonnested model selection tests of Kitamura (2000) and the nonnested Cox tests of Ramalho and Smith (2002). Although the advantage of GEL-MSC is that it facilitates choice among multiple competing models, it does have the disadvantage of not providing a framework for probabilistic statements to be made regarding the model choice (see Vuong, 1989). In contrast, the likelihood ratio testing approach of Kitamura (2000) allows probability statements about the choice of the best model in the framework of hypothesis testing.

2. MODEL SELECTION CRITERIA FOR MOMENT CONDITION MODELS

Our notation closely follows Andrews and Lu (2001). Let $g(X; \gamma)$ be the collection of moment conditions under consideration. Let b be the model selection vector that selects the elements of $\gamma \in R^p$ to be estimated, i.e., a p-dimensional vector of 0 and 1's where 1 indicates that the corresponding parameter element is to be estimated. Similarly, let c be the r-dimensional moment selection vec-

tor that selects the moment conditions in $g(\cdot) \in R^r$ to be used in the estimation. Let $\gamma_b \equiv b\% * \%\gamma$ denote the subvector of γ that is estimated and let $g_c(\cdot) \equiv c\% * \%g(\cdot;\gamma)$ denote the subvector of $g(\cdot)$ that is used in estimation, where $\% * \%$ denotes Hadamard (element-by-element) product.

In what follows, we refer to (b,c) as a pair of moment and model selection vectors. We use $|c|$ (resp. $|b|$) to denote the total number of moments (resp. parameters) selected by the pair (b,c). Furthermore, τ_c denotes the $|c|$-dimensional vector of Lagrange multipliers corresponding to the $g_c(\cdot)$ moment conditions selected by c in the construction of the GEL function described subsequently. Finally, $|c| - |b|$ is the number of overidentifying restrictions, and throughout we assume that the model is identified. This in particular requires the necessary condition that $|c| - |b| \geq 0$.

We follow Andrews and Lu (2001) in defining the following sets. Let $\mathcal{BC}$ denote the space of (b,c) vectors, which can be viewed as the "parameter space" in the moment and model selection procedure. Furthermore, define the set

$$\mathcal{BCL}^0 = \{(b,c) \in \mathcal{BC} : Eg_c(\cdot;\gamma_b) = 0, \gamma_b = \gamma\% * \%b, \quad \text{with } \gamma \in \Gamma\},$$

where $Eg_c(\cdot;\gamma_b)$ denotes the population value of the empirical moment $g_c(X;\gamma_b)$. In other words, $\mathcal{BCL}^0$ is the set of "feasible" vectors (b,c) that select only models and moments that equal zero asymptotically for some parameter vector. Finally,

$$\mathcal{MBCL}^0 = \{(b,c) \in \mathcal{BCL}^0 : |c| - |b| \geq |c^*| - |b^*| \ \forall (b^*,c^*) \in \mathcal{BCL}^0\}.$$

In short, $\mathcal{MBCL}^0$ is the set of "feasible" selection vectors (b,c) that maximize the quantity $|c| - |b|$, the number of overidentifying restrictions. Also let $\#(\mathcal{MBCL}^0)$ denote the common values of $|c| - |b|$ for all the elements of $\mathcal{MBCL}^0$.

In the rest of the paper we restrict attention to the case when $\mathcal{MBCL}^0$ is a singleton. This implies that the GEL-MSC estimator $(\hat{b}_{GMSC}, \hat{c}_{GMSC})$ defined in the next section converges to a constant and allows for the usual asymptotic distribution for the postselection parameter estimates. When $\mathcal{MBCL}^0$ is not a singleton, although the consistency result of Proposition 1 continues to hold, $(\hat{b}_{GMSC}, \hat{c}_{GMSC})$ may still be random in the limit and the asymptotic distribution for the postselection parameter estimates can be rather involved. Pötscher (1991) provides a detailed analysis of this important difference.

2.1. Generalized Empirical Likelihood-Based Model Selection Criteria

The GEL-MSC estimator, $(\hat{b}_{GMSC}, \hat{c}_{GMSC})$, minimizes GEL-based MSC over $\mathcal{BC}$. The criterion function is defined as

$$\begin{aligned} GELMSC(b,c) &= GEL_n(b,c) - h(|c| - |b|)\kappa_n \\ &\equiv 2n \min_{\gamma_b} \max_{\tau_c} Q_n(\gamma_b, \tau_c) - h(|c| - |b|)\kappa_n, \end{aligned} \tag{1}$$

where the GEL function (Newey and Smith, 2000) is defined as

$$Q_n(\gamma_b, \tau_c) = \frac{1}{n}\sum_{t=1}^{n} \rho(\tau_c' g(X_t; \gamma_b)).$$

Both $h(\cdot)$, a strictly increasing function, and the sequence κ_n are specified by the researcher. The carrier function $\rho(v)$ is a function of a scalar v that is concave on its domain $\mathcal{V}$, an open interval containing 0. It is normalized so that $\rho(0) = 0, \nabla\rho(0) = -1$, and $\nabla^2\rho(0) = -1$, where $\nabla\rho(\cdot)$ and $\nabla^2\rho(\cdot)$ correspond to the first and second derivatives of $\rho(v)$, respectively. Therefore the GEL-MSC is the usual GEL criterion function, augmented by a penalty function that varies with the number of overidentifying restrictions and also with the number of observations.

The GEL function nests several special cases of interest: when $\rho(v) = \log(1 - v)$ the GEL function corresponds to the EL function; for $\rho(v) = 1 - e^v$ it corresponds to the ET estimator; and a quadratic $\rho(\cdot)$ corresponds to the continuous updating estimator. Discussions of these estimators can be found in Newey and Smith (2000). Throughout the paper we assume that the data X_t are stationary and ergodic. The following assumptions on $h(\cdot)$ and κ_n are necessary for the consistency of the GEL-MSCs.

Assumption 1. $h(\cdot)$ is a strictly increasing function and $\kappa_n \to \infty$ as $n \to \infty$ and $\kappa_n = o(n)$.

As in Andrews and Lu (2001), examples of GEL-MSCs that satisfy Assumption 1 include analogs of the Bayesian information criterion (BIC) and Hannan and Quinn information criterion (HQIC) based on GEL, both of which use $h(x) = x$:

GEL-MSC-BIC:

$$\kappa_n = \ln n; \qquad GELMSC_{BIC}(b,c) = GEL_n(b,c) - (|c| - |b|)\ln n,$$

GEL-MSC-HQIC:

$$\kappa_n = \lambda \ln\ln n; \qquad GELMSC_{HQIC}(b,c) = GEL_n(b,c) - \lambda\cdot(|c| - |b|)\ln\ln n.$$

2.2. Consistency of GEL Model Selection Criteria

A moment and model selection estimator $(\hat{b}, \hat{c})$ is defined to be consistent if $(\hat{b}, \hat{c}) \in \mathcal{MBCL}^0$ with probability converging to 1. In the following we give a set of assumptions under which the GEL-MSC estimator $(\hat{b}_{GMSC}, \hat{c}_{GMSC})$ provides a consistent moment and model selection estimator. These assumptions allow for both random sampling and dependent data.

Assumption 2. For each $(b,c) \in \mathcal{BC}$,

(1) $\gamma_b \in \Gamma_b$, Γ_b is compact, $\tau_c \in \Lambda_c$, and Λ_c is compact and contains a neighborhood of 0.

(2) There exist interior points $\gamma^* \in \Gamma_b$ and $\tau^* \in \Lambda_c$ for each (b,c) such that

$$\min_{\gamma_b \in \Gamma_b} \max_{\tau_c \in \Lambda_c} E\rho(\tau_c' g(X_t;\gamma_b)) = E\rho(\tau_c^{*\prime} g(X_t;\gamma_b^*))$$

and $E\rho(\tau_c^{*\prime} g(X_t;\gamma_b^*)) = 0$ if $(b,c) \in \mathcal{BCL}^0$, $E\rho(\tau_c^{*\prime} g(X_t;\gamma_b^*)) > 0$ if $(b,c) \notin \mathcal{BCL}^0$.

(3) $Q_n(\hat{\gamma}_b, \hat{\tau}_c) \xrightarrow{p} E\rho(\tau_c^{*\prime} g(X_t;\gamma_b^*))$, where we denote

$$Q_n(\hat{\gamma}_b, \hat{\tau}_c) = \min_{\gamma_b \in \Gamma_b} \max_{\tau_c \in \Lambda_c} Q_n(\gamma_b, \tau_c).$$

In particular, note that for $(b,c) \in \mathcal{BCL}^0$, we can take $\tau_c^* = 0$.

The interior point condition (2) requires that the moment condition model is not too misspecified and cannot be ruled out ex ante. For example, it can be violated if $g(X_t;\gamma_b) > 0$ with probability 1 for all $\gamma_b \in \Gamma_b$. The following lemma gives a set of sufficient conditions for Assumption 2.

LEMMA 1. *Assumption (2) holds if condition (1) of Assumption 2 is satisfied and the following conditions are met for each* (b,c)*:*

(1′) $E\rho(\tau_c' g(X_t;\gamma_b))$ *is uniformly continuous over* (Γ_b, Λ_c).

(2′) For each $\gamma_b \in \Gamma_b$, $\tau_c(\gamma_b) = \arg\max_{\tau_c \in \Lambda_c} E\rho(\tau' g_c(X_t;\gamma_b))$ *is unique and is continuous in* γ_b. *The saddle point* $\gamma_b^* = \arg\min_{\gamma_b} E\rho(\tau_c(\gamma_b)' g(X_t;\gamma_b))$ *is unique.*

(3′) $\sup_{\gamma_b \in \Gamma_b, \tau_c \in \Lambda_c} |Q_n(\gamma_b, \tau_c) - E\rho(\tau_c' g(X_t;\gamma_b))| \xrightarrow{p} 0$.

A sufficient condition for the uniqueness of $\tau_c(\gamma_b)$ in condition $(2')$ of Lemma 1 is that the matrix $Eg_c(X_t;\gamma_b)g_c(X_t;\gamma_b)'$ is strictly positive definite for all (b,c) and $\gamma_b \in \Gamma_b$, because this implies that $E\rho(\tau_c' g(X_t;\gamma_b))$ is strictly convex in τ_c for each γ_b. Sufficient conditions for condition $(3')$ of Lemma 1, which also imply condition $(1')$, are as follows: (i) $g(X_t;\gamma_b)$ is uniformly continuous in γ_b; (ii) $\sup_{\gamma_b \in \Gamma_b, \tau_c \in \Lambda_c} |\rho(\tau_c' g(X_t;\gamma_b))| < \infty$; (iii) X_t are independent and identically distributed (i.i.d.). Condition (ii) is satisfied if $g(X_t;\gamma_b)$ is uniformly bounded.

Although beyond the results of this paper, much weaker conditions for uniform convergence using empirical process theory can be used to accommodate nonsmooth $g_c(X_t;\gamma_b)$ (see, e.g., Andrews, 1994). Moreover, these weaker conditions can potentially allow for more general cases such as unbounded values of $\rho(\tau_c' g(X_t;\gamma_b))$ for some realizations of X_t, which is important in the context of EL with unbounded moment functions. Consistency results under general conditions are developed by Newey and Smith (2000), who also exploit the concavity properties of the carrier function to bypass uniform convergence conditions and obtain $\sqrt{n}$-consistency in one step.

Assumption 2 ensures that with probability converging to 1, $(\hat{b}, \hat{c}) \in \mathcal{BCL}^0$. This together with the next assumption will ensure also that $(\hat{b}, \hat{c}) \in \mathcal{MBCL}^0$ with probability converging to 1.

Assumption 3. For each $(b,c) \in \mathcal{BCL}^0$, $nQ_n(\hat{\gamma}_b, \hat{\tau}_c) = O_p(1)$.

Sufficient conditions for Assumption 3 are developed in Kitamura and Stutzer (1997), Christoffersen, Hahn, and Inoue (2001), Chernozhukov and Hansen (2001), and Newey and Smith (2000), among others. Kitamura and Stutzer (1997) assume smooth moment functions. Chernozhukov and Hansen (2001) and Christoffersen et al. (2001) use empirical process theory (for nonsmooth quantile moment functions, see, e.g., Andrews, 1994). For completeness we collect some of these conditions used in the aforementioned papers in the following lemma.

LEMMA 2. *Suppose that for each* $(b,c) \in \mathcal{BCL}^0$, *there exists a unique* $\gamma_b^* \in \Gamma_b$ *such that* $Eg_c(X_t;\gamma_b) = 0$ *if and only if* $\gamma_b = \gamma_b^*$. *Assume that* $\hat{\gamma}_b \xrightarrow{p} \gamma_b^*$ *and* $\hat{\tau}_c \xrightarrow{p} 0$ *and that the following conditions are satisfied:*

($\bar{1}$*)* $\rho(\cdot)$ *is twice differentiable with bounded continuous derivatives on its domain* $\mathcal{V}$, *which includes all realizations of* $\tau_c' g_c(X_t,\gamma_b)$ *for all* $\tau_c \in \Lambda_c$ *and* $\gamma_b \in \Gamma_b$.

($\bar{2}$*) Let* $\Omega_c(\gamma_b) = Eg_c(X_t;\gamma_b)g_c(X_t;\gamma_b)'$ *be positive definite at* γ_b^*, *and for any* $\delta_n \to 0$, *suppose*

$$\sup_{|\gamma_b - \gamma_b^*| \le \delta_n} \left| \frac{1}{n}\sum_{t=1}^{n} g_c(X_t;\gamma_b)g_c(X_t;\gamma_b)' - \Omega_c(\gamma_b^*) \right| \xrightarrow{p} 0.$$

($\bar{3}$*) Asymptotic normality of moment conditions:* $1/\sqrt{n}\sum_{t=1}^{n} g_c(X_t;\gamma_b^*) = O_p(1)$.

Then Assumption 3 holds and $\sqrt{n}\hat{\tau}_c = O_p(1)$.

Note that although Lemmas 1 and 2 assume uniqueness of (γ_b^*, τ_c^*), this is not directly used in Assumptions 2 and 3. It is possible to relax these conditions to allow for nonunique γ_b^*, τ_c^*, i.e., models that are not point identified. These results are, however, beyond the scope of the paper. For correctly specified models, the results for $Q_n(\hat{\gamma}_b, \hat{\tau}_c)$ in Newey and Smith (2000) allow for unidentified moment conditions.

Given these conditions, the next proposition introduces the notion of consistency for GEL-based MSC.

Proposition 1. *Under Assumptions 1–3, we have, with probability converging to 1,* $(\hat{b},\hat{c}) \in \mathcal{MBCL}^0$ *for the pair* $(\hat{b},\hat{c}) = \arg\max_{(b,c)\in\mathcal{BC}} GELMSC(b,c)$. *In short, we say that the GEL-based* MSC *is consistent.*

Other transformations of the GEL function can also be used to form MSCs. For example, because $\log(1-x) = -x + o(x)$, the GEL-MSC may be redefined as

$$-n\log(1 - Q_n(\hat{\gamma}_b, \hat{\tau}_c)) - h(|c| - |b|)\kappa_n$$

with corresponding conditions on κ_n. In particular, in the case of exponential tilting, $Q_n(\gamma_b,\tau_c) = 1/n\sum_{t=1}^{n}(1 - e^{\tau_c' g(X_t;\gamma_b)})$, and $\log(1 - Q_n(\hat{\gamma}_b,\hat{\tau}_c))$ corresponds to the KLIC from the implied distribution to the true distribution in Kitamura and Stutzer (1997).

2.3. Time Series Data

The results of the previous section apply to both random sampling data and dependent data. For time series data, under suitable stationarity, ergodicity, and weak dependence conditions, condition (3) of Lemma (2) typically holds with (see, e.g., Andrews, 1991; Newey and West, 1987)

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} g_c(X_t;\gamma_b^*) \xrightarrow{d} N(0,S) \quad \text{where } S=\sum_{j=-\infty}^{\infty} Eg_c(X_t,\gamma_b^*)g_c(X_{t-j},\gamma_b^*).$$

For i.i.d. data, $S=\Omega_c(\gamma_b^*)$ and the GEL estimator $\hat{\gamma}_b$ is as efficient as the optimally weighted GMM estimator for $(b,c)\in\mathcal{MBCL}^0$. For time series data, although the GEL-MSC in the previous section is still consistent by Proposition 1, the estimator $\hat{\gamma}_b$ is typically less efficient than the optimally weighted GMM.

To achieve efficiency with dependent data, the blocking methods of Kitamura and Stutzer (1997) (in the special case of ET) and Smith (1997) can be used to smooth the observations in constructing the GEL function. These authors define the blockwise GEL objective function as

$$\bar{Q}_n(\gamma_b,\tau_c)=\frac{1}{n}\sum_{t=1}^{n}\rho(\tau_c'\hat{g}_c(X_t;\gamma_b)) \quad \text{where}$$

$$\hat{g}_c(X_t;\gamma_b)=\frac{1}{2K+1}\sum_{k=-K}^{K} g_c(X_{t-k};\gamma_b)$$

and the blocks are constructed so that $K\to\infty$ and $K/\sqrt{n}\to 0$. In what follows, we briefly discuss the modifications needed when $\bar{Q}_n(\gamma_b,\tau_c)$ is used in place of $Q_n(\gamma_b,\tau_c)$ in constructing consistent GEL-MSCs. We only outline the results based on Kitamura and Stutzer (1997) and Smith (1997) and refer the reader to these papers for the complete set of stationary and weak dependence conditions and other regularity conditions that validate the blockwise GEL approach.

When $(b,c)\in\mathcal{BCL}^0$, $(\hat{\gamma}_b,\hat{\tau}_c)\xrightarrow{p}(\gamma_b^*,0)$ and $\bar{Q}_n(\hat{\gamma}_b,\hat{\tau}_c)\xrightarrow{p}0$. On the other hand, for misspecified models in which $(b,c)\notin\mathcal{BCL}^0$, $(\hat{\gamma}_b,\hat{\tau}_c)\xrightarrow{p}(\gamma_b^*,\tau_c^*)$ and $\bar{Q}_n(\hat{\gamma}_b,\hat{\tau}_c)\xrightarrow{p}E\rho(\tau_c^{*'}g_c(X_t;\gamma_b^*))>0$. Furthermore, when $(b,c)\in\mathcal{BCL}^0$

$$\frac{2n}{2K+1}\bar{Q}_n(\hat{\gamma}_b,\hat{\tau}_c)=O_p(1).$$

Following the logic of Proposition 1, consistent MSC can be defined as before by minimizing $n\bar{Q}_n(\hat{\gamma}_b,\hat{\tau}_c)-h(|c|-|b|)\kappa_n$ over (b,c), where the condition on κ_n is now modified to $\kappa_n/K\to\infty$ and $\kappa_n/n\to 0$.

2.4. Testing Procedure

Given the general consistency result of GEL-based MSC, we also describe the GEL analogs of two algorithms proposed in Andrews (1999) and Andrews and

Lu (2001) to choose (b,c) consistently. In the following discussion we will focus on $Q_n(\gamma_b,\tau_c)$ rather than $\bar{Q}_n(\gamma_b,\tau_c)$.

2.4.1. Downward testing procedure. Andrews and Lu (2001) define the downward-testing model selection procedure as follows. Starting with vectors $(b,c) \in \mathcal{BC}$ for which $|c| - |b|$ (the number of overidentifying restrictions) is the largest, perform tests (described in detail subsequently) with progressively smaller $|c| - |b|$ (therefore the name "downward" testing) until a test is found that cannot reject the null hypothesis that the moment conditions considered are all correct for the given model. (Note that for each value of $|c| - |b|$, tests are carried out for each (b,c) in $\mathcal{BC}$ with this value of $|c| - |b|$.) Let $\hat{k}_{DT}$ denote the number of overidentifying restrictions (i.e., $|c| - |b|$) for this first test found to not reject the null. Given $\hat{k}_{DT}$, the downward testing estimator $(\hat{b}_{DT},\hat{c}_{DT})$ is the vector that maximizes $GEL_n(b,c)$ over $(b,c) \in \mathcal{BC}$ with $|c| - |b| = \hat{k}_{DT}$.

More formally, consider the GEL statistic: $GEL_n(b,c) = 2nQ_n(\hat{\gamma}_b,\hat{\tau}_c)$. Under Assumptions 2 and 3, if the moment conditions are correctly specified (in the sense that $\tau_c^* = 0$ for the limit GEL problem $\min_{\gamma_b}\max_{\tau_c}[E\rho(\tau_c' g(X_t;\gamma_b))]$), then $GEL_n(b,c) = O_p(1)$.[2]

The downward-testing procedure looks for the first acceptance among $(b,c) \in \mathcal{BC}$ of the test whose rejection region is defined by

$$GEL_n(b,c) \geq \eta_{n,k} = \chi_k^2(\alpha_n),$$

where $\chi_k^2(\alpha_n)$ denotes the $(1-\alpha_n)$th quantile of the chi-squared distribution with $k = |c| - |b|$ degrees of freedom. The following consistency result can be shown for the downward-testing estimators $(\hat{b}_{DT},\hat{c}_{DT})$, which is analogous to Theorem 2 in Andrews and Lu (2001).

Proposition 2. *If the sequence of critical values satisfies for each k*

$$\eta_{n,k} \to \infty \quad \textit{and} \quad \eta_{n,k} = o(n) \quad \textit{as } n \to \infty,$$

then under Assumptions 2 and 3, $P((\hat{b}_{DT},\hat{c}_{DT}) \in \mathcal{MBCL}^0) \xrightarrow{p} 1$.

2.4.2. Upward testing procedure. The GELs can also be applied to the upward-testing procedure described in Andrews (1999). Starting with vectors $(b,c) \in \mathcal{BC}$ that have the smallest number of overidentifying restrictions $|c| - |b|$, we perform tests (analogous to those described for the downward-testing procedure previously) with progressively more overidentifying restrictions (i.e., larger $|c| - |b|$; therefore the name "upward testing") until we find that all tests with the same value of $|c| - |b|$ reject the null hypothesis that the moment conditions considered are all correct. Let $\hat{k}_{UT}$ denote the largest value such that for all $k \leq \hat{k}_{UT}$, there is at least one $(b,c) \in \mathcal{BC}$ with $|c| - |b| = k$ for which the null hypothesis is not rejected. Given $\hat{k}_{UT}$, we take the upward

testing estimator $(\hat{b}_{UT}, \hat{c}_{UT})$ to be the vector that minimizes $GEL_n(b,c)$ over $(b,c) \in \mathcal{BC}$ with $|c| - |b| = \hat{k}_{UT}$.

It is necessarily true that the upward testing procedure described here will never select a pair (b,c) with more overidentifying restrictions than the downward testing procedure; i.e.

$$|\hat{b}_{UT}| - |\hat{c}_{UT}| \leq |\hat{b}_{DT}| - |\hat{c}_{DT}|. \tag{2}$$

To avoid selecting a pair (b,c) with too few overidentification conditions, an additional assumption (as in Andrews, 1999) is made to ensure consistency of $\hat{b}_{UT}$ and $\hat{c}_{UT}$.

Assumption 4. For each $(b,c) \in \mathcal{BC}$ such that $k \equiv |c| - |b| < \#(\mathcal{MBCL}^0)$, there exists (b,c) with $|c| - |b| = k$ for which $(b,c) \in \mathcal{BCL}^0$.

Without this condition, the inequality (2) may hold strictly, even asymptotically. Note that this additional condition can be ensured by proper choice of the parameter space $\mathcal{BC}$ for the selection vector (b,c). Under this additional condition, we state the following proposition.

Proposition 3. *With probability converging to 1,* $(\hat{b}_{UT}, \hat{c}_{UT}) \in \mathcal{MBCL}^0$.

2.5. Analogy with Parametric Likelihood Model Selection Procedure

Andrews (1999) shows that the *J*-statistic-based MSC is analogous to standard MSC (such as the BIC, Akaike information criterion [AIC], and HQIC) often employed in parametric model selection procedures. When we use GEL to formulate the MSC, this analogy is very transparent because, in this case, an explicit likelihood- (or information-) based rationale also underlies the moment selection procedure, just as in the fully parametric case.

Andrews (1999) notes that his *J*-statistic MSC is analogous to the parametric MSC in the sense that, among correct models, this criterion would choose the more tightly specified model. Equation (6.6) in Andrews (1999) shows an equivalence result between the problem of maximizing the number of moment conditions (i.e., minimizing the number of excluded moment conditions) and minimizing the number of parameters, among correctly specified models. In this section, we show an analogous equivalence for GEL-based MSCs.

Under correct specification GEL is asymptotically equivalent to GMM estimation using the optimal weighting matrix. The use of GEL-based MSC also provides a transparent proof of this equivalence result by avoiding the issues associated with choosing the optimal weighting matrix in GMM estimation, which arise when considering the *J*-statistic.

Following Andrews (1999), we simplify notation by assuming that all the models under consideration are correctly specified, and we focus on the moment selection problem (involving the moment selection vector c and the associated Lagrange multipliers τ_c). Therefore, in the discussion that follows, we

let $b \equiv \vec{1}$, the vector whose elements are all 1, and $\gamma_b = \gamma$ throughout, and we assume that $g_c(\cdot)$ is sufficient for identification of γ. Our goal is to show the equivalence between

$$GEL_{1n} = 2 \min_{\gamma} \max_{\tau_c} \left[\sum_{t=1}^{n} \rho(\tau_c' g_c(X_t; \gamma)) \right], \tag{3}$$

where $g_c(\cdot)$ is the subvector of $g(\cdot)$ selected by c, and

$$GEL_{2n} = 2 \min_{\gamma, \mu} \max_{\tau_c, \tau_{-c}} \left[\sum_{t=1}^{n} \rho(\tau_c' g_c(X_t; \gamma) + \tau_{-c}'(g_{-c}(X_t; \gamma) - \mu)) \right], \tag{4}$$

where $g_{-c}(\cdot)$ is the subvector of the totality of moment conditions $g(\cdot)$ that are excluded by the selection vector c. Here μ is of dimension $r - |c|$, where r is the total number of moment conditions under consideration. This equivalence is analogous to equation (6.6) in Andrews (1999) and implies that the moment selection problem can alternatively be viewed as a model (i.e., parameter) selection problem, with the augmented parameter vector $(\gamma, \mu)'$.

The equivalence of (3) and (4) is easy to demonstrate; indeed, let $(\tilde{\gamma}, \tilde{\tau}_c)$ solve (3), i.e., satisfy the first-order conditions

$$2 \sum_{t=1}^{n} g_c(X_t; \tilde{\gamma}) \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) = 0,$$

$$2 \sum_{t=1}^{n} \frac{\tilde{\tau}_c' \partial g_c(X_t; \tilde{\gamma})}{\partial \gamma} \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) = 0.$$

Then it follows that $(\gamma = \tilde{\gamma}, \tau_c = \tilde{\tau}_c, \tau_{-c} = 0)$ and

$$\mu = \left(\sum_{t=1}^{n} g_{-c}(X_t; \tilde{\gamma}) \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) \right) \Big/ \left(\sum_{t=1}^{n} \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) \right)$$

solves the problem in (4). Indeed, one can easily verify that the first-order conditions for problem (4), which can be written as

$$\frac{\partial}{\partial \tau_c} GEL_{2n} = 2 \sum_{t=1}^{n} g_c(X_t; \tilde{\gamma}) \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) = 0,$$

$$\frac{\partial}{\partial \gamma} GEL_{2n} = 2 \sum_{t=1}^{n} \tilde{\tau}_c' \frac{\partial g_c(X_t; \tilde{\gamma})}{\partial \gamma} \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) = 0,$$

$$\frac{\partial}{\partial \tau_{-c}} GEL_{2n} = 2 \sum_{t=1}^{n} (g_{-c}(X_t; \tilde{\gamma}) - \mu) \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) = 0,$$

$$\frac{\partial}{\partial \mu} GEL_{2n} = 2 \tau_{-c} \sum_{t=1}^{n} \nabla \rho(\tilde{\tau}_c' g_c(X_t; \tilde{\gamma})) = 0,$$

are necessarily satisfied at these parameter values. It is also immediately obvious that at these parameter values the two GEL functions are identical:

$$GEL_{1n}(\tilde{\gamma}, \tilde{\tau}_c) = GEL_{2n}(\tilde{\gamma}, \mu, \tilde{\tau}_c, 0),$$

which is analogous to equation (6) in Andrews (1999). Thus the analogy between generalized empirical likelihood-based moment and model selection procedures and Andrews' *J*-statistic based procedures is complete. The use of GEL-based MSC allows us to generalize the likelihood-based rationale underlying the usual MSC for parametric models to semiparametric models in which the data-generating process is only partially specified via population moment restrictions.

3. MONTE CARLO EXPERIMENTS

In this section we report the results from a simple Monte Carlo study designed to compare MSC based on the *J*-statistic as proposed by Andrews (1999) and Andrews and Lu (2001) and also on two special cases of the GEL statistic: empirical likelihood and exponential tilting. Formally, these criteria are written as

$$MSCJ_n(b,c) = \min_{\gamma_b} n g_{nc}(\gamma_b)' W_n g_{nc}(\gamma_b) - h(|c| - |b|)\kappa_n,$$

$$g_{nc}(\gamma_b) = \frac{1}{n}\sum_{t=1}^{n} g_c(X_t; \gamma_b),$$

$$MSCEL_n(b,c) = \min_{\gamma_b} \max_{\tau} 2\sum_{t=1}^{n} \log(1 - \tau' g_c(X_t; \gamma_b)) - h(|c| - |b|)\kappa_n,$$

$$MSCET_n(b,c) = \min_{\gamma_b} \max_{\tau} 2\sum_{t=1}^{n} (1 - \exp(\tau' g_c(X_t; \gamma_b))) - h(|c| - |b|)\kappa_n,$$

using the same notation as in the previous sections.

Appropriate choices of the $h(\cdot)$ function and the sequence of constants, κ_n, deliver the BIC, AIC, and HQIC MSC. We also consider the choice of $h(\cdot)$ as the identity mapping and sequence of constants as $\kappa_n = \sqrt{n}$, which we refer to as root *n* information criterion (RNIC).

The model is specified by the set of equations

$$y_t = 1 + x_t + 0.5u_t(1 + \alpha|z_t|),$$

$$x_t = \eta_t + 0.5u_t, \qquad z_t = \eta_t + 0.5\phi_t, \qquad f_t = \eta_t + 0.3u_t,$$

where u_t, η_t, and ϕ_t are all independently distributed $N(0,1)$ random variables, truncated at -2 and 2. Both z_t and f_t are candidate instruments. By considering alternative values for the coefficient α the analysis can accommodate homoskedastic and heteroskedastic error structures. To this end, values of zero and a small positive constant are considered. The fact that $E[\, f_t u_t] \neq 0$ implies that

moment conditions constructed from the instrument f_t are invalid. Moment conditions are constructed from the following five possible instrument groups.

M1. constant, $\cos(z) + \sin(z)$.

M2. constant, z, $\cos(z) + \sin(z)$, $\cos(z)$.

M3. constant, $\cos(z) + \sin(z)$, $\sin(f)$.

M4. constant, z, $\cos(z) + \sin(z)$, $\cos(z)$, $\sin(f)$.

M5. constant, $\cos(z) + \sin(z)$, $\cos(f)$, $\sin(f)$.

The econometrician is assumed to know that the M1 moment conditions are valid and seeks to determine the verity of the remaining moment conditions for estimation. In the preceding notation, M2 instruments are the true (b^0, c^0), M1 instruments are other consistent (b, c), and the remaining instrument groups contain inconsistent (b, c). Therefore, this model that we use for the Monte Carlo experiments satisfies the condition that $\mathcal{MBCL}^0$ is a singleton. This is a case in which the limiting distribution of the postselection (for consistent MSCs) estimator is known to be the same as if (b^0, c^0) are picked a priori (see, e.g., Pötscher, 1991).

In the subsequent tables, these three groupings of instruments will be referenced by the abbreviations OC, Truth, and IC, respectively. The $\sin(\cdot)$ and $\cos(\cdot)$ functions are utilized as a convenient way to generate instruments. Following the suggestions in Andrews and Lu (2001), we have chosen the moment conditions in these models so that (i) there is a noticeable difference in efficiency between the estimators that use all the correct moment conditions and the estimators that use only those moment conditions that are known to be correct and (ii) there are noticeable biases in the estimators that use incorrect moment conditions. This setup allows for gains to be exploited from a good moment selection procedure. It is clear that the exercise can be generalized to allow also for "model selection" over sets of possible regressors.

Following Andrews and Lu (2001), we assess the relative performance of these MSC by comparing the probability with which the three MSCs select the true (b^0, c^0); other consistent (b, c); and inconsistent (b, c). MSC that have both a high probability of selecting the true model and a low probability of selecting inconsistent models are preferred. The performance of postselection estimators is assessed by comparison of the bias and root mean squared errors (RMSEs) of the estimated slope coefficient of each model. The rejection rates for a 5% *t*-test that each of these parameter estimates is equal to the true value of unity are also computed. These statistics are reported for estimation based on each of the five instrument sets and for each of the four MSC, to give a total of nine postselection model results. Because there are three estimation methods—GMM, EL, and ET—we report postselection results for a total of 27 models. Results based on the proposed EL- and ET-based MSC relative to the GMM-based approach are of most interest. The results based on each of the five pos-

sible instrument groups are presented for comparison, with the infeasible estimator resulting from the use of M2 giving the "ideal" benchmark. The results are based on 500 repetitions for four sample sizes, $N = 50$, 250, 500, and 1,000.

Table 1 details the probabilities of selecting the true, other consistent, and inconsistent models under assumption of a homoskedastic error structure ($\alpha = 0$). For each estimation method, results are collected by each of the four proposed penalty functions (BIC, AIC, HQIC, and RNIC). Considering the BIC and $\sqrt{N}$ criteria, it is clear that the *J*-MSC outperforms both the MSCEL and MSCET for the smaller sample sizes. The probability of selecting the true model is higher by 5% and the probability of selecting a misspecified model lower by up to 10%. As the sample size increases the discrepancy between the *J*-MSC and MSCET vanishes, whereas the MSCEL has slightly higher probability of selecting an inconsistent model. Under the AIC criterion, the MSCET outperforms both *J*-MSC and MSCEL at all sample sizes. The gains are of the order of 10%, though they are somewhat smaller when compared to MSCEL at larger sample

TABLE 1. Selection probabilities

	J-statistic			Empirical likelihood			Exponential tilting		
N	OC	Truth	IC	OC	Truth	IC	OC	Truth	IC
				BIC criterion					
50	0.010	0.678	0.312	0.002	0.610	0.388	0.000	0.582	0.478
250	0.004	0.982	0.014	0.000	0.958	0.042	0.000	0.982	0.018
500	0.000	1.000	0.000	0.000	0.974	0.026	0.000	1.000	0.000
1,000	0.000	1.000	0.000	0.000	0.982	0.018	0.000	1.000	0.000
				AIC criterion					
50	0.068	0.698	0.234	0.006	0.630	0.364	0.002	0.750	0.246
250	0.152	0.842	0.006	0.032	0.914	0.054	0.028	0.958	0.014
500	0.162	0.838	0.000	0.026	0.926	0.048	0.028	0.972	0.000
1,000	0.144	0.856	0.000	0.014	0.950	0.036	0.002	0.980	0.000
				HQIC criterion					
50	0.026	0.708	0.266	0.012	0.668	0.320	0.002	0.654	0.244
250	0.024	0.966	0.010	0.004	0.948	0.048	0.004	0.978	0.018
500	0.028	0.972	0.000	0.000	0.968	0.032	0.000	1.000	0.000
1,000	0.020	0.980	0.000	0.000	0.974	0.026	0.000	1.000	0.000
				RNIC criterion					
50	0.014	0.698	0.288	0.006	0.630	0.364	0.000	0.618	0.382
250	0.000	0.986	0.014	0.000	0.960	0.040	0.000	0.982	0.018
500	0.000	1.000	0.000	0.000	0.978	0.022	0.000	1.000	0.000
1,000	0.000	1.000	0.000	0.000	0.994	0.006	0.000	1.000	0.000

sizes. The inconsistency of the AIC selection procedure is immediate from the *J*-MSC results (though interestingly, in the context of this model, the inconsistency of the AIC criterion seems to be small under EL- and ET-based methods and AIC appears able to distinguish the correct model with high probability). Finally, for HQIC, *J*-MSC again performs better in the smallest sample size, whereas MSCET is marginally better as N increases. For all models, we note that the MSC appear to perform reasonably well for sample sizes above 250.

Table 2 presents the bias, RMSEs, and rejection rates of the postselection estimates of the model's slope coefficient for the two smallest sample sizes. To clarify notation, for each estimation method (GMM, EL, and ET), the results based on each of the five instrument groups are labeled M1–M5, whereas those postselection results arising from the four MSC are labeled by the corresponding penalty term: i.e., BIC, AIC, HQIC, and RNIC.

Considering the results obtained under GMM, it is immediate that misspecification can lead to poor postselection results. If estimates are based on any of the incorrectly specified models (M3–M5) then the bias is some 10 times greater than the infeasible estimator (M2), with a corresponding deterioration in the RMSEs. For such models, the rejection rate in small sample size ($N = 50$) for a 5% *t*-test that the slope coefficient is equal to the true value of unity is likely to be rejected over 60% of the time. Thus, misspecification can clearly lead to erroneous conclusions. In contrast, for the MSC, we note that the performance of the postselection estimators is much closer to the infeasible estimator. The bias is approximately three times that under M2, whereas the RMSEs are marginally higher. Correspondingly, the rejection rates of a 5% *t*-test are reduced for each of the MSC relative to misspecified models to about 20%. Of the four selection criteria, the AIC seems to perform best in small sample sizes ($N = 50$), even though it is theoretically inconsistent. This does not appear too surprising, given the finite-sample bias and size distortion when $N = 50$ and the fact that the selection probabilities for AIC compare very favorably to other MSCs in this small sample size of $N = 50$. In addition, the post-AIC estimators are still consistent, and the distortion in the sampling distribution might not be significant for the model we consider and for small sample sizes. For larger sample sizes, however, AIC clearly does not perform quite as well as other consistent MSCs, even though the discrepancy appears marginal. As the sample size increases to 250, the differences between results based on the true model and those based on each of the four selection criteria are remarkably small. The biases are comparable, whereas the RMSEs and rejection rates are slightly higher.

For the EL- and ET-based results, the same broad patterns of results are observed. However, comparison of these results to those obtained under GMM suggests that the former have smaller bias and comparable RMSEs for the smallest sample size, though a slightly higher rejection rate under ET. For the $N = 250$ sample, ET seems to perform better than EL. The bias is half as much and the RMSEs somewhat smaller. Relative to GMM, ET seems to perform marginally better.

TABLE 2. Postselection results 1

	$N = 50$			$N = 250$		
	Bias	RMSE	Rej. Rate	Bias	RMSE	Rej. Rate
			GMM			
M1	−0.006	0.115	0.078	0.000	0.046	0.056
M2	0.011	0.091	0.148	0.004	0.036	0.068
M3	0.139	0.152	0.674	0.137	0.139	0.998
M4	0.128	0.143	0.620	0.119	0.119	0.982
M5	0.130	0.146	0.616	0.122	0.122	0.990
BIC	0.032	0.110	0.256	0.005	0.039	0.084
AIC	0.021	0.111	0.224	0.003	0.041	0.080
HQIC	0.027	0.111	0.224	0.004	0.039	0.084
RNIC	0.031	0.111	0.256	0.005	0.038	0.084
			Empirical likelihood			
M1	−0.006	0.115	0.078	0.000	0.046	0.056
M2	−0.006	0.099	0.152	0.001	0.036	0.072
M3	0.128	0.144	0.626	0.125	0.128	0.992
M4	0.097	0.124	0.470	0.098	0.102	0.932
M5	0.116	0.137	0.536	0.107	0.111	0.958
BIC	0.028	0.109	0.226	0.006	0.042	0.108
AIC	0.020	0.108	0.236	0.008	0.045	0.124
HQIC	0.022	0.110	0.240	0.007	0.044	0.116
RNIC	0.025	0.111	0.252	0.005	0.042	0.108
			Exponential tilting			
M1	−0.006	0.115	0.078	0.000	0.046	0.056
M2	−0.004	0.113	0.150	0.001	0.036	0.062
M3	0.132	0.148	0.644	0.131	0.133	0.996
M4	0.104	0.135	0.514	0.096	0.100	0.910
M5	0.117	0.138	0.536	0.113	0.116	0.976
BIC	0.029	0.118	0.300	0.003	0.039	0.080
AIC	0.020	0.113	0.254	0.025	0.039	0.082
HQIC	0.024	0.115	0.270	0.003	0.039	0.082
RNIC	0.027	0.117	0.284	0.003	0.039	0.080

Finally, Table 3 presents an identical set of results to those in Table 2 but for the sample sizes $N = 500$ and 1,000. Consistent with the results of Table 1, as N increases from 250 to 1,000 the bias is essentially eliminated for GMM- and ET-based models and the RMSEs fall at a rate consistent with the increase in sample size. For both larger sample sizes the ET-based results are marginally better than the GMM-based results, whereas the EL-based results are somewhat worse. For ET and GMM the BIC, HQIC, and RNIC MSC nearly always

TABLE 3. Postselection results 2

	$N = 500$			$N = 1{,}000$		
	Bias	RMSE	Rej. Rate	Bias	RMSE	Rej. Rate
			GMM			
M1	0.000	0.033	0.052	0.000	0.025	0.068
M2	0.001	0.025	0.056	0.000	0.019	0.066
M3	0.135	0.137	1.000	0.135	0.136	1.000
M4	0.118	0.120	1.000	0.118	0.118	1.000
M5	0.121	0.123	1.000	0.122	0.122	1.000
BIC	0.001	0.025	0.056	0.000	0.019	0.066
AIC	0.002	0.030	0.068	0.000	0.022	0.088
HQIC	0.000	0.026	0.058	0.000	0.020	0.068
RNIC	0.001	0.025	0.056	0.000	0.019	0.068
			Empirical likelihood			
M1	0.000	0.033	0.052	0.000	0.025	0.068
M2	−0.001	0.025	0.060	−0.001	0.019	0.060
M3	0.123	0.125	1.000	0.124	0.124	1.000
M4	0.096	0.099	0.992	0.097	0.098	1.000
M5	0.106	0.108	1.000	0.107	0.107	1.000
BIC	0.002	0.031	0.084	0.001	0.025	0.076
AIC	0.004	0.036	0.106	0.004	0.030	0.092
HQIC	0.003	0.032	0.090	0.002	0.027	0.084
RNIC	0.002	0.029	0.080	0.000	0.020	0.064
			Exponential tilting			
M1	0.000	0.033	0.052	0.000	0.025	0.068
M2	−0.001	0.025	0.058	−0.001	0.019	0.060
M3	0.129	0.131	1.000	0.129	0.130	1.000
M4	0.095	0.097	0.994	0.095	0.096	1.000
M5	0.111	0.113	1.000	0.112	0.113	1.000
BIC	−0.001	0.025	0.058	−0.001	0.019	0.060
AIC	−0.001	0.026	0.062	0.000	0.020	0.062
HQIC	−0.001	0.025	0.058	−0.001	0.019	0.060
RNIC	−0.001	0.025	0.058	−0.001	0.019	0.060

select the true model for the largest sample size. As expected, for large N the rejection rates go to 1 for misspecified models and are in the neighborhood of 5% for the infeasible model and all MSC-based selection models.

We also study the performance of selection probabilities and postselection estimators in a heteroskedastic version of the preceding model. The Monte Carlo results for a heteroskedastic case are not reported given their similarity to the homoskedastic results, but they can be obtained at the Web address www.princeton.edu/~doubleh.

4. CONCLUSIONS

This paper, following Andrews (1999) and Andrews and Lu (2001), proposes GEL-based MSC for unconditional moment-based models. The MSC seek to minimize the GEL-statistic modified by a penalty function that rewards use of additional correct moment conditions for a given number of parameters and penalizes less tightly specified models for a given number of moment conditions.

The GEL-based criteria have an information-theoretic interpretation even if all models are incorrectly specified. If there is at least one model that is correctly specified then the GEL-MSC chooses the most parsimoniously specified model among correctly specified models with probability converging to 1. If all models are misspecified, the proposed MSC choose the model that minimizes the penalty-augmented GEL-statistic. Thus, in the case of EL, the consistent MSC chooses the model that is closest to the population density in KLIC distance and also the most parsimonious in the number of parameters.[3]

The usefulness of the GEL-based MSC was considered in a simple Monte Carlo study for two special cases of GEL: empirical likelihood and exponential tilting. Whereas in small sample sizes, *J*-MSC performed well relative to the EL- and ET-based MSC for a range of postselection statistics, in larger samples the ET-based MSC performed marginally better than the *J*-MSC, with some improvements over the EL-based MSC. Although these results are specific to the example we studied, they suggest that GEL-based MSC can be a useful alternative to *J*-MSC.

NOTES

1. For detailed definitions see Andrews and Lu (2001). We closely follow their notation.

2. For random sampling data typically $GEL_n(b,c)$, as defined in (1), converges in distribution to $\chi^2_{|c|-|b|}$ under correct specification, although for consistent model selection we only need $GEL_n(b,c) = O_p(1)$.

3. Using GEL to form MSC also has the advantage of not having to choose a weighting matrix. Another benefit is that GEL criterion functions remain invariant to certain normalizations of moment conditions. We thank a referee for raising this point with us.

REFERENCES

Ahn, H., Y. Kitamura, & G. Tripathi (2001) Empirical Likelihood-Based Inference in Conditional Moment Restriction Models. Working paper, Department of Economics, University of Wisconsin.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.

Andrews, D.W. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–858.

Andrews, D. (1994) Empirical Process Methods in Econometrics. In R. Engle & D. McFadden (eds.), *Handbook of Econometrics*, Vol. 4, pp. 2248–2292. Amsterdam: North-Holland.

Andrews, D. (1999) Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 67, 543–564.

Andrews, D. & B. Lu (2001) Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* 101, 123–164.

Chernozhukov, V. & C. Hansen (2001) An IV Model of Quantile Treatment Effects. Working paper, Department of Economics, MIT.

Christoffersen, P.F., J. Hahn, & A. Inoue (2001) Testing and comparing value at risk measures. *Journal of Empirical Finance* 8, 325–342.

Hansen, L., J. Heaton, & A. Yaron (1996) Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14, 262–280.

Imbens, G., R. Spady, & P. Johnson (1998) Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–357.

Kitamura, Y. (1997) Empirical likelihood methods with weakly dependent processes. *Annals of Statistics* 25, 2084–2102.

Kitamura, Y. (2000) Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood. Working paper, Department of Economics, University of Wisconsin.

Kitamura, Y. & M. Stutzer (1997) An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65, 861–874.

Newey, W. & D. McFadden (1994) Large sample estimation and hypothesis testing. In R. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2113–2241. Amsterdam: North-Holland.

Newey, W. & R. Smith (2000) Asymptotic Bias and Equivalence of GMM and GEL Estimators. Working paper 01/517, University of Bristol.

Newey, W. & K.D. West (1987) A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.

Pötscher, B. (1991) Effects of model selection on inference. *Econometric Theory* 7, 163–185.

Qin, J. & J. Lawless (1994) Empirical likelihood and general estimating equations. *Annals of Statistics* 22, 300–325.

Ramalho, J.J. & R.J. Smith (2002) Generalized empirical likelihood non-nested tests. *Journal of Econometrics* 102, 1–28.

Smith, R.J. (1997) Alternative semiparametric likelihood approaches to generalised method of moments estimation. *Economics Journal* 107, 503–519.

Vuong, Q. (1989) Likelihood-ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333.

APPENDIX: PROOFS

Proof of Lemma 1. Existence of (γ_b^*, τ_c^*) in condition (2) of Assumption 2 is ensured by the continuity of $E\rho(\tau_c(\gamma_b)'g(X_t, \gamma_b))$ and the compactness of the parameter space. First consider condition (3) of Assumption 2. Define

$$\hat{\tau}_c(\gamma_b) = \text{argmax}_{\tau_c \in \Lambda_c} Q_n(\gamma_b, \tau_c).$$

Using all three conditions, standard arguments as in Amemiya (1985) and Newey and McFadden (1994) adjusted for the dependence of the objective function on γ_b can be used to show that $\sup_{\gamma_b \in \Gamma_b} |\hat{\tau}_c(\gamma_b) - \tau_c(\gamma_b)| = o_p(1)$. Therefore by the definition that $\hat{\gamma}_b = \text{argmin}_{\gamma_b \in \Gamma_b} Q_n(\hat{\tau}_c(\gamma_b), \gamma_b)$, to show $\hat{\gamma}_b - \gamma_b = o_p(1)$, it suffices to show that

$$\sup_{\gamma_b \in \Gamma_b} |Q_n(\hat{\tau}_c(\gamma_b), \gamma_b) - E\rho(\tau_c(\gamma_b)'g_c(X_t, \gamma_b))| = o_p(1).$$

This can be split into two parts. The first part is

$$\sup_{\gamma_b \in \Gamma_b} |Q_n(\hat{\tau}_c(\gamma_b), \gamma_b) - E\rho(\hat{\tau}_c(\gamma_b)' g_c(X_t, \gamma_b))| = o_p(1),$$

and the second part is $\sup_{\gamma_b \in \Gamma_b} |E\rho(\hat{\tau}_c(\gamma_b)' g_c(X_t, \gamma_b)) - E\rho(\tau_c(\gamma_b)' g_c(X_t, \gamma_b))| = o_p(1)$. The first part follows from condition (3′), and the second part follows from condition (1′) and uniform convergence of $\hat{\tau}_c(\gamma_b)$ to $\tau_c(\gamma_b)$.

Consider part (2) next. For $\bar{\gamma}_b$ such that $Eg_c(X_t; \bar{\gamma}_b) = 0$,

$$\frac{\partial}{\partial \tau_c} E\rho(\tau_c' g_c(X_t; \bar{\gamma}_b))\bigg|_{\tau_c = 0} = \nabla\rho(0) Eg_c(X_t; \bar{\gamma}_b) = 0.$$

By concavity of $\rho(\cdot)$, $E\rho(\tau_c' g_c(X_i; \bar{\gamma}_b))$ achieves a maximum value of 0 when $\tau_c = 0$. Therefore for all $(b, c) \in \mathcal{BCL}^0$, by the uniqueness assumption in condition (2′), $E\rho(\tau_c' g_c(X_i; \gamma_b))$ achieves a value of 0 at the unique saddle point $(\gamma_b^*, \tau_c^* = 0)$. On the other hand, for $\bar{\gamma}_b$ such that $Eg_c(X_i; \bar{\gamma}_b) \neq 0$,

$$\frac{\partial}{\partial \tau_c} E\rho(\tau_c' g_c(X_t; \bar{\gamma}_b))\bigg|_{\tau_c = 0} = \nabla\rho(0) Eg_c(X_t; \bar{\gamma}_b) \neq 0,$$

so that $\tau_c(\bar{\gamma}_b) \neq 0$ by the concavity of $\rho(\cdot)$. Hence by the uniqueness assumption (2′),

$$E\rho(\tau_c(\bar{\gamma}_b) g_c(X_t; \bar{\gamma}_b)) > 0.$$

Therefore for $(b, c) \in \mathcal{BC}$, but $\notin \mathcal{BCL}^0$, $E\rho(\tau_c^{*\prime} g_c(X_t; \gamma_b^*)) > 0$, as part (2) requires. ■

Proof of Lemma 2. The arguments follow those in Kitamura and Stutzer (1997), Christoffersen et al. (2001), Chernozhukov and Hansen (2001), and Newey and Smith (2000). We summarize the key steps here. First define $\hat{\tau}_c^* = \text{argmax}_{\tau_c \in \Lambda_c} \sum_{t=1}^n \rho(\tau_c' g_c(X_t; \gamma_b^*))$. Using conditions ($\bar{1}$) and ($\bar{2}$) to Taylor-expand the first-order condition

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n g_c(X_t; \gamma_b^*) \nabla\rho(\hat{\tau}_c^{*\prime} g_c(X_t; \gamma_b^*)) = 0$$

around $\tau_c = 0$, we obtain that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n g_c(X_t; \gamma_b^*) + \Omega_c(\gamma_b^*) \sqrt{n} \hat{\tau}_c^* + o_p(\sqrt{n} \hat{\tau}_c^*)$$

$$\Rightarrow \sqrt{n} \hat{\tau}_c^* = \Omega_c(\gamma_b^*)^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^n g_c(X_t; \gamma_b^*) + o_p(1) = O_p(1),$$

where the last equality follows from conditions ($\bar{2}$) and ($\bar{3}$). Then a quadratic expansion shows

$$\sum_{t=1}^n \rho(\hat{\tau}_c^{*\prime} g_c(X_t; \gamma_b^*)) = \sqrt{n} \hat{\tau}_c^* \frac{1}{\sqrt{n}} \sum_{t=1}^n g_c(X_t; \gamma_b^*) + (\sqrt{n} \hat{\tau}_c^*)' \Omega_c(\gamma_b^*)(\sqrt{n} \hat{\tau}_c^*) + o_p(n \hat{\tau}_c^{*2})$$

$$= O_p(1).$$

Then, for arbitrary h,

$$O_p(1) = \sum_{t=1}^{n} \rho(\hat{\tau}_c^{*'} g_c(X_t, \gamma_b^*)) \geq \sum_{t=1}^{n} \rho\left(\frac{h}{\sqrt{n}} g_c(X_t, \hat{\gamma}_b)\right)$$

$$= \frac{h}{\sqrt{n}} \sum_{t=1}^{n} g_c(X_t; \hat{\gamma}_b) + h' \Omega_c(\gamma_b^*) h + o_p(h^2),$$

which implies $1/\sqrt{n}\sum_{t=1}^{n} g_c(X_t; \hat{\gamma}_b) = O_p(1)$. Next, Taylor-expand the first-order condition for $\hat{\tau}_c$:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g_c(X_t; \hat{\gamma}_b) \nabla \rho(\hat{\tau}_c' g_c(X_t; \hat{\gamma}_b)) = 0$$

to obtain that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g_c(X_t; \hat{\gamma}_b) + \Omega_c(\gamma_b^*) \sqrt{n} \hat{\tau}_c + o_p(\sqrt{n} \hat{\tau}_c)$$

$$\Rightarrow \sqrt{n} \hat{\tau}_c = \Omega_c(\gamma_b^*)^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} g_c(X_t; \hat{\gamma}_b) + o_p(1) = O_p(1).$$

Finally, using $1/\sqrt{n}\sum_{t=1}^{n} g_c(X_t; \hat{\gamma}_b) = O_p(1)$ and $\sqrt{n}\hat{\tau}_c = O_p(1)$ in a Taylor expansion, we obtain

$$\sum_{t=1}^{n} \rho(\hat{\tau}_c' g_c(X_t; \hat{\gamma}_b)) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} g_c(X_t, \hat{\gamma}_b) \sqrt{n} \hat{\tau}_c + \sqrt{n} \hat{\tau}_c' \Omega_c(\gamma_b^*) = O_p(1).$$

As noted in Newey and Smith (2000), the preceding result does not require the complete set of conditions for $\sqrt{n}$-consistency and asymptotic normality of $\hat{\gamma}_b$. The conditions of the lemma can potentially be modified to allow for the cases when γ_b^* is not uniquely identified. ■

Proof of Proposition 1. The proof is very similar to Andrews and Lu (2001). Because by Assumption 2 condition (1) the domain of Λ_c includes 0 as an interior point, by the saddle point definition of (γ_b^*, τ_c^*), for each $(b, c) \in \mathcal{BC}$: $E\rho(\tau_c^{*'} g(X_t; \gamma_b^*)) \geq 0$. Take $(b, c) \in \mathcal{BC}$ but $\notin \mathcal{BCL}^0$. By the uniqueness Assumption 2 condition (2), $E\rho(\tau_c^{*'} g(X_t; \gamma_b^*)) > 0$. Then by Assumption 2 condition (3),

$$Q_n(\hat{\gamma}_b, \hat{\tau}_c) \xrightarrow{p} E\rho(\tau_c^{*'} g(X_t; \gamma_b^*)) > 0.$$

So, using Assumption 1 that $\kappa_n/n \to 0$,

$$\frac{1}{2n} GELMSC_n(b, c) \xrightarrow{p} E\rho(\tau_c^{*'} g(X_i; \gamma_b^*)) > 0.$$

On the other hand, if $(b,c) \in \mathcal{BCL}^0$, $E\rho(\tau_c' g(X_i;\gamma_b))$ achieves a value of 0 at the unique saddle point $(\gamma_b^*, \tau_c^* = 0)$. Therefore, again using $\kappa_n/n \to 0$,

$$\frac{1}{2n} GELMSC_n(b,c) \xrightarrow{p} 0.$$

Hence, the preceding two equations imply that $(\hat{b},\hat{c}) \in \mathcal{BCL}^0$ with probability converging to 1.

On the other hand, for all $(b,c) \in \mathcal{BCL}^0$, $Q_n(\hat{\gamma}_b, \hat{\tau}_c) = O_p(1)$. But for $|c_1| - |b_1| < |c_2| - |b_2|$ (i.e., the pair (b_2,c_2) has more overidentifying restrictions than the pair (b_1,c_1)), such that both pairs are in $\mathcal{BCL}^0$, $(h(|c_1| - |b_1|) - h(|c_2| - |b_2|))\kappa_n \to -\infty$.

Therefore, with probability converging to 1, $GELMSC_n(b_2,c_2) < GELMSC_n(b_1,c_1)$, namely, that $(\hat{b},\hat{c}) \in \mathcal{MBCL}^0$ with probability converging to 1. ■

Proof of Proposition 2. For any $(b,c) \in \mathcal{BC}$ but $\notin \mathcal{BCL}^0$, the proof of Proposition 1 has shown that

$$GEL_n(b,c)/\eta_{n,|c|-|b|} \xrightarrow{p} \infty$$

because in this case, $GEL_n(b,c)$ is $O_p(n)$.

Thus $\hat{k}_{DT} \leq \#(\mathcal{MBCL}^0)$ *w.p.* $\to 1$. On the other hand, for $(b,c) \in \mathcal{BCL}^0$, under Assumption 3,

$$GEL_n(b,c) < \eta_{n,|c|-|b|} \quad w.p. \to 1.$$

In consequence, $\hat{k}_{D,T} = \#(\mathcal{MBCL}^0)$ *w.p.* $\to 1$, and hence $(\hat{b}_{DT}, \hat{c}_{DT}) \in \mathcal{MBCL}^0$. ■

Proof of Proposition 3. For the same reason as in the previous proof, we see that $\hat{k} = |\hat{c}_{UT}| - |\hat{b}_{UT}| \leq \#(\mathcal{MBCL}^0)$ *w.p.* $\to 1$. On the other hand, Assumption 4 implies that each $k = |c| - |b| < \#(\mathcal{MBCL}^0)$; we can find corresponding b_k and c_k such that $(b_k,c_k) \in \mathcal{BCL}^0$, under which it is necessary that

$$GEL_n(b_k,c_k) < \eta_{n,|c_k|-|b_k|} \quad w.p. \to 1.$$

Consequently, with probability tending to 1, $\hat{k}_{UT} = |\hat{c}_{UT}| - |\hat{b}_{UT}| = \#(\mathcal{MBCL}^0)$ and $(\hat{b}_{UT}, \hat{c}_{UT}) \in \mathcal{MBCL}^0$. ■